

This is a postprint version of the following published document:

Drugman, T. & Dutoit, T. (2013). *Advances in Nonlinear Speech Processing: 6th International Conference, NOLISP 2013, Mons, Belgium, June 19-21, 2013. Proceedings.* (pp. 9-11). (Lecture Notes in Computer Science; 7911). Springer Berlin Heidelberg. DOI: http://dx.doi.org/10.1007/978-3-642-38847-7_2

NMF-Based Spectral Analysis for Acoustic Event Classification Tasks

Jimmy Ludeña-Choez^{1,2} and Ascensión Gallardo-Antolín¹

¹ Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid,
Avda. de la Universidad 30, 28911 - Leganés (Madrid), Spain

² Facultad de Ingenierías, Universidad Católica San Pablo, Arequipa, Perú
{jimmy,gallardo}@tsc.uc3m.es

Abstract. In this paper, we propose a new front-end for Acoustic Event Classification tasks (AEC). First, we study the spectral contents of different acoustic events by applying Non-Negative Matrix Factorization (NMF) on their spectral magnitude and compare them with the structure of speech spectra. Second, from the findings of this study, we propose a new parameterization for AEC, which is an extension of the conventional Mel Frequency Cepstrum Coefficients (MFCC) and is based on the high pass filtering of acoustic event spectra. Also, the influence of different frequency scales on the classification rate of the whole system is studied. The evaluation of the proposed features for AEC shows that relative error reductions about 12% at segment level and about 11% at target event level with respect to the conventional MFCC are achieved.

Keywords: Acoustic Event Classification, Non-Negative Matrix Factorization, Auditory Filterbank.

1 Introduction

In recent years, the problem of automatically detecting and classifying acoustic non-speech events has attracted the attention of numerous researchers. Although speech is the most informative acoustic event, other kind of sounds (such as laughs, coughs, keyboard typing, etc.) can give relevant cues about the human presence and activity in a certain scenario (for example, in an office room). This information could be used in different applications, mainly in those with perceptually aware interfaces such as smart-rooms [1]. Additionally, acoustic event detection and classification systems, can be used as a pre-processing stage for automatic speech recognition (ASR) in such way that this kind of sounds can be removed prior to the recognition process increasing its robustness. In this paper, we focus on acoustic event classification (AEC).

A design of a suitable feature extraction process for AEC is an important issue. Several front-ends have been proposed in the literature, some of them based on short-term features, such as Mel-Frequency Cepstral Coefficients (MFCC) [1], [2], [3], [4], log filterbank energies [3], Perceptual Linear Prediction (PLP) [5], log-energy, spectral flux, fundamental entropy and zero-crossing rate [1].

Other approaches are based on the application of different temporal integration techniques over these short-term features [6], [7].

However, as pointed in [3] these features are not necessarily the more appropriate for AEC tasks because they have been design according to the spectral characteristics of speech which are quite different from the spectral structure of acoustic events. To deal with this issue, in [3], it is proposed a boosted feature selection method to construct a more suitable parameterization for AEC.

In this work, we follow a different approach. First, we study the spectral characteristics of different acoustic events by applying Non-Negative Matrix Factorization (NMF) on their spectral magnitude and compare them with the structure of speech spectra. As NMF provides a way to decompose a signal into a convex combination of non-negative building blocks (called Spectral Basis Vectors, SBV) by minimizing a cost function, the resulting SBVs carry the information about the most relevant spectral components of each acoustic event. Second, from the findings of this study, we propose a new parameterization for AEC, which is an extension of the conventional MFCC and is based on the high pass filtering of acoustic event spectra. Also, the influence of different frequency scales (Mel, ERB, Bark and linear) on the classification rate of the whole system is studied.

This paper is organized as follows: Section 2 introduces the mathematical background of NMF. In Section 3 we present the spectral analysis of acoustic events using NMF. Section 4 is devoted to the explanation to our proposed parameterization and Section 5 describes the experiments and results to end with some conclusions and ideas for future work in Section 6.

2 Non-negative Matrix Factorization (NMF)

Given a matrix $V \in \mathbb{R}_+^{F \times T}$, where each column is a data vector, NMF approximates it as a product of two matrices of nonnegative low rank W and H , such that

$$V \approx WH \quad (1)$$

where $W \in \mathbb{R}_+^{F \times K}$ and $H \in \mathbb{R}_+^{K \times T}$ and normally $K \leq \min(F, T)$. This way, each column of V can be written as a linear combination of the K basis vectors (columns of W), weighted with the coefficients of activation or gain located in the corresponding column of H . NMF can be seen as a dimensionality reduction of data vectors from an F -dimensional space to the K -dimensional space. This is possible if the columns of W uncover the latent structure in the data [8]. The factorization is achieved by an iterative minimization of a given cost function as, for example, the Euclidean distance or the generalized Kullbak Leibler (KL) divergence,

$$D_{\text{KL}}(V \| WH) = \sum_{ij} \left(V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - (V - WH)_{ij} \right) \quad (2)$$

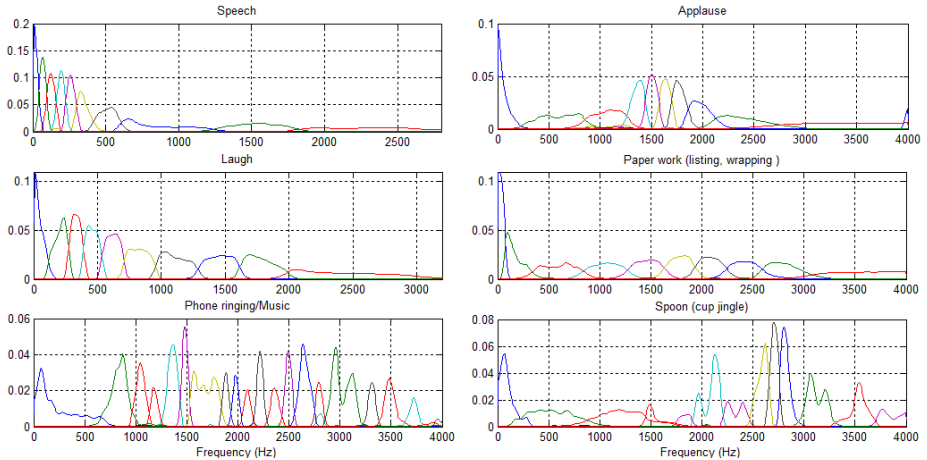


Fig. 1. Spectral Basis Vectors (SBVs) for speech and different acoustic events

In this work, we consider the KL divergence because it has been recently used with good results in speech processing tasks, such as speech enhancement and denoising for ASR tasks [9] [10] or feature extraction [11]. In order to find a local optimum value for the KL divergence between V and (WH) , an iterative scheme with multiplicative update rules can be used as proposed in [8] and stated in (3),

$$W \leftarrow W \otimes \frac{V H^T}{1 H^T} \quad H \leftarrow H \otimes \frac{W^T V}{W^T 1} \quad (3)$$

where 1 is a matrix of size V , whose elements are all ones and the multiplications \otimes and divisions are component wise operations. NMF algorithm produces a sparse representation of the data, reducing the redundancy.

3 NMF-Based Spectral Analysis of Acoustic Events

In order to gain insight into the spectral content of the different Acoustic Events (AEs) considered, a NMF-based spectral analysis of each of these acoustic classes have been carried out.

For doing this, for a given AE, NMF is applied to the short-term spectrum magnitude of a subset of the audio files belonging to this particular class. The spectral basis vectors of this AE, W_e , are obtained minimizing the KL divergence between the magnitude spectra $|V_e|$ and their corresponding factored matrices $W_e H_e$ using the learning rules in (3). Note that the matrix W_e contains the SBVs that can be seen as the building blocks which represent each AE, as it is verified that $|V_e| \approx W_e H_e$.

The SBVs of five different non-speech sounds (applause, laugh, paper work, phone ringing and spoon cup jingle) are represented in Figure 1. The SBVs of

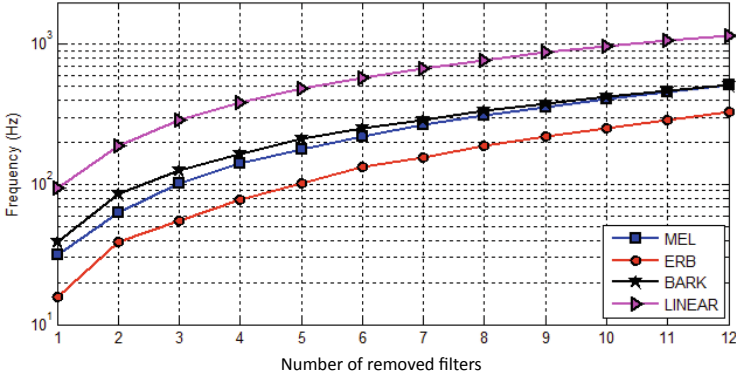


Fig. 2. Upper frequency of the stopband vs. number of removed filters

speech are also represented for comparison purposes. In all cases, 10 SBVs were obtained by applying NMF to the short-term spectrum magnitude computed over 20 ms windows with a frameshift of 10 ms. From this figure, the following observations can be extracted:

- The spectral content of the AEs are very different each other and with respect to speech. In fact, while the spectral components of speech are concentrated in low frequencies, the non-speech sounds present, in general, a relevant spectral content in medium-high frequencies.
- In all cases, low frequency components are presented to a greater or lesser extent, so this part of the spectrum seems not to very discriminative when comparing different types of AEs (including speech).
- Comparing the SBVs of the non-speech sounds, it can be observed that large differences can be found in the medium-high part of the spectrum, suggesting that these frequency bands are more suitable (or at least, they can not be negligible) than the lower part of the spectrum for discriminating between different acoustic events.

4 Parameterization Derived from the High-Pass Filtering of the Acoustic Event Spectrum

The analysis of the SBVs of the different acoustic events shown in Section 3 motivated us to derive a modified version of the conventional MFCC in which the special relevance of the medium-high frequencies of the spectrum is taking into account. This can be accomplished by filtering the short-term spectrum of the signal (using the appropriate high-pass filter) prior to the application of the auditory filterbank (in the case of MFCC, a mel-scaled triangular filterbank). However, in this work, we adopt a straightforward method which consists of modifying the auditory filterbank by means of the explicit removal of a certain number of the filters placed on the low frequency bands of the spectrum.

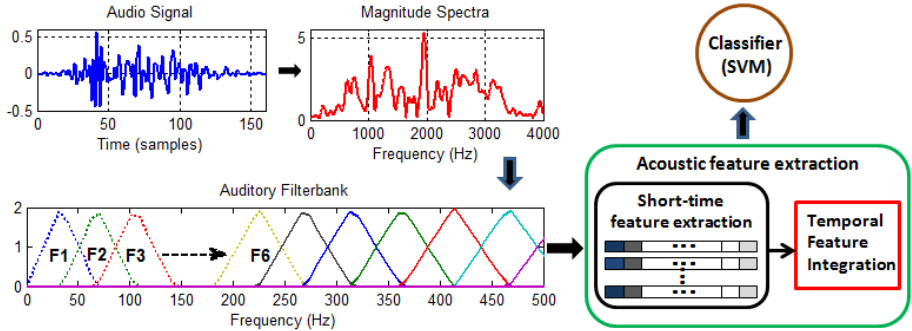


Fig. 3. Block diagram of the proposed method

In addition, in order to analyse the influence of the filter positions for AEC, several well-known frequency scales are considered: Mel, ERB, Bark and linear.

In Figure 2 it can be observed the upper frequency of the complete stopband as a function of the number of removed filters in the auditory filterbank for the four scales considered.

Once the speech spectrum is filtered following the procedure previously described and the remaining log filterbank energies are computed, a Discrete Cosine Transform is applied over them as in the case of the conventional MFCC yielding to a set of cepstral coefficients. Finally, it is applied a temporal feature integration technique which consists of dividing the sequence of cepstral coefficients into segments and computing the statistics of these parameters (in this case, mean, standard deviation and skewness) over each segment. These segment-based parameters are the input to the acoustic event classifier, which is based on Support Vector Machines (SVM). This process is summarized in Figure 3.

5 Experiments

5.1 Database and Experimental Protocol

The database used for the experiments consists of a total of 2,114 instances of target events belonging to 12 different acoustic classes: applause, coughing, chair moving, door knock, door slam, keyboard typing, laughter, paper wrapping, phone ringing, steps, spoon/cup jingle and key jingle. The composition of the whole database was intended to be similar to the one used in [3]. Audio files were obtained from different sources: websites, the FBK-Irst database [12] and the UPC-TALP database [13]. The speech sounds used for the computation of the speech SBVs shown in Figure 1 were extracted from the ShATR database [14].

Since this database is too small to achieve reliable classification results, we have used a 6-fold cross validation to artificially extend it, averaging the results afterwards. Specifically, we have split the database into six disjoint balanced

Table 1. Average classification rate [%] (segment) for different frequency scales

Param.	Scale	Number of Eliminated Filters												
		Base.	1	2	3	4	5	6	7	8	9	10	11	12
CC	MEL	75.10	77.47	77.66	77.58	77.63	78.16	76.95	78.11	76.87	76.12	77.23	77.23	76.10
	ERB	74.02	74.74	75.95	77.38	77.43	77.53	76.81	76.77	77.09	76.66	77.76	76.90	76.71
	BARK	74.30	77.39	77.27	77.68	76.96	77.31	76.27	77.43	76.91	76.72	77.11	76.77	76.59
	LINEAR	77.29	77.30	77.62	76.84	77.26	75.52	75.33	74.96	73.88	74.43	73.36	73.22	71.83
CC+ Δ CC	MEL	77.57	79.43	79.45	79.22	79.36	79.07	79.20	79.55	79.41	78.47	77.81	78.77	78.55
	ERB	76.51	77.57	78.80	79.14	79.42	78.69	79.22	79.13	79.04	78.74	79.20	78.79	78.97
	BARK	77.58	78.98	79.32	78.64	78.65	78.33	78.62	79.25	78.86	78.77	78.03	78.08	78.56
	LINEAR	79.09	80.39	79.94	78.16	78.88	78.82	78.15	76.64	76.54	76.27	76.54	76.42	75.54

groups. One different group is kept for testing in each fold, while the remainder are used for training.

The AEC system is based on a one-against-one SVM with RBF kernel and a majority voting scheme for the final decision [7]. For each one of these experiments, a 5-fold cross validation was used for computing the optimal values of RBF kernel parameters.

5.2 Results

For the baseline experiments, 12 cepstral coefficients were extracted every 10 ms using a Hamming analysis window of 20 ms long and an auditory filterbank composed of 40 spectral bands. Four different frequency scales were considered: Mel (yielding to the conventional MFCC), ERB, Bark and linear. Also, the log-energy of each frame and the first derivatives (where indicated) were computed and added to the cepstral coefficients. The final feature vectors consisted of the statistics of these short-term parameters (mean, standard deviation and skewness) computed over segments of 2 s length with overlap of 1 s.

Table 1 and Table 2 show, respectively, the results achieved in terms of the average classification rate at segment level (percentage of segments correctly classified) and at target event level (percentage of target events correctly classified) by varying the number of eliminated low frequency bands in the auditory filterbank. Results for the baseline systems (when no frequency bands are eliminated) are also included. Both tables contain the classification rates for the four frequency scales considered (Mel, ERB, Bark and linear) and for two different set of acoustic parameters (CC: cepstral coefficients + log-energy and CC+ Δ CC: cepstral coefficients + log-energy + its derivatives).

As can be observed for the CC parameterization, the performance of the Mel, ERB and Bark scales are quite similar, being the Mel scale slightly better. The behaviour with respect to the elimination of low frequency bands follows the same trends for the three scales. In all cases, the high pass filtering of the acoustic event spectrum outperforms the baseline: for the Mel scale, the best performance is achieved when the number of eliminated filters varies from 3 to 7, for the ERB

Table 2. Average classification rate [%] (target event) for different frequency scales

Param.	Scale	Number of Eliminated Filters												
		Base.	1	2	3	4	5	6	7	8	9	10	11	12
CC	MEL	81.07	82.28	82.04	82.42	82.42	81.89	81.31	83.20	81.27	80.78	80.69	81.75	79.72
	ERB	79.43	80.73	81.46	82.09	82.57	82.52	82.71	82.42	82.28	81.46	83.29	81.51	80.73
	BARK	80.83	81.94	82.47	82.33	80.83	81.07	80.98	81.84	80.73	81.07	80.98	81.55	80.98
	LINEAR	82.04	80.98	81.12	80.49	80.44	79.19	78.51	77.89	76.29	77.16	77.02	76.24	74.70
CC+ Δ CC	MEL	81.41	82.62	83.39	83.58	83.49	83.15	82.38	82.71	82.81	80.06	81.12	81.55	81.22
	ERB	80.73	80.98	82.18	82.67	83.24	82.62	82.76	81.89	82.04	81.80	82.71	81.75	82.57
	BARK	81.84	82.76	82.71	81.41	82.62	81.84	82.04	82.09	81.55	81.80	81.22	81.41	81.22
	LINEAR	82.81	82.38	82.42	81.60	81.36	80.78	80.35	79.33	79.24	79.04	79.38	78.71	77.16

scale, from 3 to 10 and for the Bark scale, from 2 to 7. From Figure 2, it can be seen that these ranges of eliminated filters roughly correspond to a stopband from 0 Hz to 100-275 Hz. The linear scale outperforms the classification rates achieved with the other scales in the baseline experiment (when no frequency bands are removed). However, no further improvements are obtained when low frequency filters are eliminated from the auditory filterbank. This can be explained for the higher bandwidth of the low frequency filters in the linear scale with respect to the other scales.

In summary, when using CC parameters, the best performance is obtained with the Mel scale when the seven first low frequency filters are not considered in the cepstral coefficients computation. In this case, the difference in performance with respect to the baseline is statistically significant at 95% confidence level and the relative error reduction with respect to the respective baseline is around 12% at segment level and around 11% at target event level.

Similar observations can be drawn for the CC+ Δ CC parameterization: best results are obtained when low frequencies (below 100-275 Hz) are not considered in the feature extraction process. When comparing to CC, it can be observed that CC+ Δ CC achieves improvements about 1% absolute over CC, However, these differences are not statistically significant.

6 Conclusion

In this paper, we have presented a new parameterization method for acoustic event classification tasks, motivated by the study of the spectral characteristics of non-speech sounds. First, we have analysed the spectral contents of different acoustic events by applying NMF on their spectral magnitude and compared them with the structure of speech spectra, concluding that medium and high frequencies are specially important for the discrimination between non-speech sounds. Second, from the findings of this study, we have proposed a new front-end for AEC, which is an extension of the MFCC parameterization and is based on the high pass filtering of acoustic event spectra. We have compared the proposed features to the conventional MFCC for an AEC task, obtaining relative error reductions about 12% at segment level and about 11% at target event level.

For future work, we plan to use feature selection techniques for automatically determining the most discriminative frequency bands for AEC. Other future lines include the unsupervised learning of auditory filter banks by means of NMF.

Acknowledgments. This work has been partially supported by the Spanish Government grants TSI-020110-2009-103, IPT-120000-2010-24 and TEC2011-26807. Financial support from the Fundación Carolina and Universidad Católica San Pablo, Arequipa (Jimmy Ludeña-Choez) is thankfully acknowledged.

References

1. Temko, A., Nadeu, C.: Classification of acoustic events using SVM-based clustering schemes. *Pattern Recognition* 39, 684–694 (2006)
2. Zieger, C.: An HMM based system for acoustic event detection. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) *RT 2007 and CLEAR 2007*. LNCS, vol. 4625, pp. 338–344. Springer, Heidelberg (2008)
3. Zhuang, X., Zhou, X., Hasegawa-Johnson, M.A., Huang, T.S.: Real-world acoustic event detection. *Pattern Recognition Letters* 31, 1543–1551 (2010)
4. Kwangyoun, K., Hanseok, K.: Hierarchical approach for abnormal acoustic event classification in an elevator. In: *IEEE Int. Conf. AVSS*, pp. 89–94 (2011)
5. Portelo, J., Bugalho, M., Trancoso, I., Neto, J., Abad, A., Serralheiro, A.: Non speech audio event detection. In: *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1973–1976 (2009)
6. Meng, A., Ahrendt, P., Larsen, J.: Temporal feature integration for music genre classification. *IEEE Trans. on Audio, Speech, and Language Processing* 15, 1654–1664 (2007)
7. Mejía-Navarrete, D., Gallardo-Antolín, A., Peláez, C., Valverde, F.: Feature extraction assesment for an acoustic-event classification task using the entropy triangle. In: *Interspeech*, pp. 309–312 (2011)
8. Lee, D., Seung, H.: Algorithms for non-negative matrix factorization. *Nature* 401, 788–791 (1999)
9. Wilson, K., Raj, B., Smaragdis, P., Divakaran, A.: Speech denoising using nonnegative matrix factorization with priors. In: *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4029–4032 (2008)
10. Ludeña-Choez, J., Gallardo-Antolín, A.: Speech denoising using non-negative matrix factorization with kullback-leibler divergence and sparseness constraints. In: Torre Toledano, D., Ortega Giménez, A., Teixeira, A., González Rodríguez, J., Hernández Gómez, L., San Segundo Hernández, R., Ramos Castro, D. (eds.) *IberSPEECH 2012*. CCIS, vol. 328, pp. 207–216. Springer, Heidelberg (2012)
11. Schuller, B., Weninger, F., Wollmer, M.: Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In: *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4562–4565 (2010)
12. FBK-Irst database of isolated meeting-room acoustic events, ELRA Catalog no. S0296
13. UPC-TALP database of isolated meeting-room acoustic events, ELRA Catalog no. S0268
14. The ShATR multiple simultaneous speaker corpus,
<http://www.dcs.shef.ac.uk/spandh/projects/shatrweb/index.html>